Received: 18 July 2020 | Revised: 15 April 2022 | Accepted: 26 April 2022

DOI: 10.1002/sim.9431

RESEARCH ARTICLE

Statistics in Medicine WILEY

The impact of methodological choices when developing predictive models using urinary metabolite data

Nikolas Krstic[1] | **Kevin Multani**[1,2] | **David S. Wishart**[3] | **Tom Blydt-Hansen**[4] | **Gabriela V. Cohen Freue**[1]

[1]Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada

[2]Department of Physics, Stanford University, Stanford, California, USA

[3]Departments of Computing Science and Biological Sciences, University of Alberta, Edmonton, Alberta, Canada

[4]Department of Pediatrics, University of British Columbia, Vancouver, British Columbia, Canada

Correspondence
Gabriela V. Cohen Freue, Department of Statistics, University of British Columbia, 2207 Main Mall, Vancouver, BC, Canada V6T 1Z4.
Email: gcohen@stat.ubc.ca

Funding information
Discovery Grant, Natural Sciences and Engineering Research Council of Canada; Undergraduate Student Research Award, Natural Sciences and Engineering Research Council of Canada; Kidney Foundation of Canada, Grant/Award Number: KFOC170003; CHIR, Grant/Award Number: #274755; Undergraduate Student Research Award; Natural Sciences and Engineering Research Council of Canada; Canada Foundation for Innovation

Abstract

The continuous evolution of metabolomics over the past two decades has stimulated the search for metabolic biomarkers of many diseases. Metabolomic data measured from urinary samples can provide rich information of the biological events triggered by organ rejection in pediatric kidney transplant recipients. With additional validation, metabolic markers can be used to build clinically useful diagnostic tools. However, there are many methodological steps ranging from data processing to modeling that can influence the performance of the resulting metabolomic classifiers. In this study we focus on the comparison of various classification methods that can handle the complex structure of metabolomic data, including regularized classifiers, partial least squares discriminant analysis, and nonlinear classification models. We also examine the effectiveness of a physiological normalization technique widely used in the clinical and biochemical literature but not extensively analyzed and compared in urine metabolomic studies. While the main objective of this work is to interrogate metabolomic data of pediatric kidney transplant recipients to improve the diagnosis of T cell-mediated rejection (TCMR), we also analyze three independent datasets from other disease conditions to investigate the generalizability of our findings.

KEYWORDS

machine learning, predictive modeling, sample quality, T cell-mediated rejection, urinary metabolites

1 | INTRODUCTION

Notable technical and analytical advancements in metabolomics have contributed to the development of diverse predictive models to diagnose or guide the treatment of various medical conditions,[1-7] including renal transplant rejection.[8-11] Many transplant patients are at risk of developing T cell-mediated rejection (TCMR) as their T cells "recognize" donor-specific molecules as antigens and evoke an immune response to eliminate the foreign transplant tissue.[12,13] The development of noninvasive urinary tests resulting from metabolomics discoveries to clinically diagnose and monitor TCMR episodes can greatly improve patient care.[14-18] In the last decade, steadily advancing biological, technological, and experimental resources have been complemented with the development of improved

statistical and computational methodologies to distill the rich information contained in metabolomic datasets.[19-22] However, the impact of different methodological approaches on classification performance is still not fully understood.

Many data processing techniques are specific to the technology used to generate the metabolomic data [eg, nuclear magnetic resonance (NMR) or liquid chromatography mass spectrometry (LC-MS)], as well as the type of biological sample used (eg, tissue, serum, or urine). Several previously published studies have extensively evaluated the effect of different data processing approaches on the identification of metabolomic markers and their classification accuracy for different types of metabolomic data.[19-21] However, most of these results are based on only one particular statistical method to construct a classifier (usually PLS, PCA, or SVM). Unfortunately, a detailed comparison of different methods to construct metabolomic classifiers from pre-processed data has not yet been done. To address this shortcoming, we have chosen to focus our analysis on a comparison of different types of classification methods, including regularized, nonlinear and tree-based methods, with regard to classifiers' performance. We also compare the effects regarding sensitivity to sample quality, and some widely used preprocessing steps for LC-MS metabolomic data.[20,21]

Among the many urine normalization techniques proposed and studied, normalization to creatinine remains a preferred choice to scale or adjust urine metabolite concentrations in the clinical and biochemical literature.[19,23] We note that the term normalization can have different meanings in different disciplines. To follow the terminology used in related papers, the term normalization in this study refers to scaling the metabolite concentrations relative to creatinine concentration. The aim of this physiological normalization is to control for the effect of dilution of urine samples, especially with regard to metabolite concentrations. Despite the simplicity of using a ratio to scale each metabolite relative to creatinine, the effectiveness of this technique depends on physiological and numerical assumptions that may not always be true for the given data.[19,23-25] Numerically, this normalization procedure assumes that each metabolite concentration decreases linearly with the concentration of creatinine, potentially at a different rate in the treatment compared to the control groups. Nevertheless, this normalization is widely used in metabolomic studies making a thorough analysis of its impact on a variety of classification methods essential.

Transformation is another important analytical step used in metabolomic studies to reduce the heteroscedasticity or skewness usually observed in urinary metabolite concentrations. While many techniques have been previously compared using both NMR and LC-MS data, the natural logarithmic transformation remains the most commonly used technique in the analysis of metabolomic data.[19,20] Thus, we focus our analysis on the assessment of the logarithmic transformation effects on metabolite concentrations and on the performance of various classifiers compared in our study.

Our work has been motivated by the search of better classifiers of TCMR based on targeted LC-MS urinary metabolomic data and medical history data that can improve the diagnosis of allograft rejection, to be used to monitor treatment using noninvasive urinary tests. As a result, we used this targeted urinary LC-MS data on TCMR to compare various classification methods and to understand the effects of some classical preprocessing techniques. We then further explore the generalizability of these results with three additional urinary LC-MS metabolomic datasets of other conditions.

2 | METHODS

2.1 | TCMR data overview

Since 2002, a prospective cohort study conducted by the transplant manitoba pediatric kidney program (TMPKP) has been collecting medical history and metabolite data from pediatric renal transplant patients (under 19 years of age) at Winnipeg's Health Sciences Centre.[26] The institutional review board of the University of Manitoba granted approval of the study (IRB file: HS14740; H2002:070), while subjects that participated in the study provided assent with parental informed consent. The data from this prospective cohort was used by Blydt-Hansen et al to build a predictive model of renal rejection,[11] which inspired this work.

The dataset includes the urinary samples from 59 pediatric patients that were collected during the period of 2002 to 2013. Each patient has between 2 and 12 samples, resulting in an initial total of 396 urinary samples. Collection of urine samples typically occurred on the morning of the scheduled kidney biopsies of the patients. Thus, in general, the samples were collected at varying time points after transplantation. The biopsies were classified using the Banff 2007 classification criteria[27] and by the same pathologist (who was blinded to the measured urine metabolites). Further details of urine sample collection, biopsy procedures, and urinalysis can be found in the study by Blydt-Hansen et al.[11]

In this article, we removed samples that contained BK virus (BKV) nephropathy, recurrent glomerular disease, or urinary tract infection (UTI). These conditions may affect metabolite concentrations and thus impair our ability to accurately classify TCMR.[28-30] Additionally, we removed samples if either the presence of donor-specific antibodies (DSA) could not be determined or a complete set of Banff scores (i, t, g, or ptc) were not available. After applying these filters, the total number of samples present in our dataset was 363. Medical history data are also available for every patient. These variables include sex, race, age at transplant, original renal disease, donor type (deceased or living donor), age of donor, years on dialysis and estimated glomerular filtration rate (eGFR).

A total of 132 metabolites were identified and absolutely quantified from every urine sample using a targeted LC-MS assay. Further information on the full set of urinary metabolites measured (as well as their standard concentrations) can be found in the publication by Bouatra et al.[31] Metabolite concentrations were measured in μmol/L. Each metabolite has a different limit of detection (LOD) and values below that limit were discarded, thus generating left-censored data. We do not know their exact values, only that they are below the metabolite-specific LODs. Using this protocol, approximately 42% of the values among the metabolite data are left-censored. To impute the missing values, we substituted the left-censored values with half of the metabolite-specific LOD ($0.5 \times \text{LOD}$) for every metabolite. While this is a common technique in many metabolomic studies, a comparison of other imputation techniques has been previously performed.[21]

We used the statistical programming language R (Version 3.6.0) to perform all of our analyses in this study.[32]

2.2 | TCMR classification

Using Banff t- and i-scores to classify TCMR in kidney transplant pediatric patients is challenging due to the qualitative nature of scoring and imprecision of inter-rater reliability. This is particularly true of scores in the lower range (~1), when inflammation is milder and where the risk of downgrading (0) or upgrading (2) is more subjective. We present one possible approach that involves using a decision rule based on the summations of the Banff t- and i-scores to perform a binary classification:

- **IF** (t-score + i-score) < 3, **THEN** classify sample as "non-TCMR"
- **ELSE IF** (t-score + i-score) > = 4, **THEN** classify sample as "TCMR"

This classification scheme is still comparable to the Banff 2007 classification criteria.[27] The "non-TCMR" and "TCMR" classes capture biopsies that would be classified similarly under the Banff criteria. However, this decision rule also captures a few biopsies in the "non-TCMR" and "TCMR" classes that would fall in the upper and lower ranges of "borderline" classification under the Banff criteria. Since there continues to be an ongoing discussion of whether the Banff criteria's "borderline" category may be too wide (Becker et al; Nankivell et al) we believe that the chosen scheme reflects, in most cases, two distinct phenotypes.[33,34]

Furthermore, this classification excludes potential borderline samples with a score sum of 3 from further analysis, resulting in a final total of 334 samples from 59 patients. Borderline cases are notoriously difficult to correctly classify, where inter-rater variability may lead to a diagnosis from mild inflammation to definitive TCMR.[33] Thus, although borderline cases are clinically relevant, our analysis focuses on the binary classification of clear non-TCMR vs clear TCMR samples.

2.3 | Exploratory principal component analysis

Before performing any modeling, we explored the metabolite data using principal component analysis (PCA) to identify any abnormalities in the data (eg, batch effects or artifacts). The dimensionality reduction offered by PCA can also provide a clearer understanding of the variability within this high dimensional metabolite dataset. We first distinguished the patients into one of two categories: (1) "Patients with at least one TCMR sample" and (2) "Patients with non-TCMR samples." Among patients in the first category, we randomly selected a single TCMR sample for each of these patients. Then, we uniquely matched each TCMR sample with a sample from a patient in the second category by the post-transplant time (in months) when the TCMR sample was collected. The matching was performed such that the sample times between the pairs of samples were as close as possible. This is quite similar to a case-control matching procedure. The metabolite data from all of the samples were standardized prior to conducting PCA.

2.4 | Data preprocessing

Prior to modeling, we generated multiple different metabolite datasets using different data preprocessing steps to examine their impact on prediction performance. For every dataset, we conducted data filtering as outlined in Sections 2.1 and 2.2. We also performed different combinations of two possible data preprocessing steps. The first data preprocessing step involved the normalization of the metabolite data by creatinine, in which each metabolite concentration within a sample was divided by the corresponding creatinine value. The second data preprocessing step involved a natural logarithmic transformation of the raw or ratio (creatinine normalized) concentrations.

To examine the effect of sample dilution on the prediction performance of each model, we removed samples with a measured specific gravity below several different thresholds. These thresholds were selected within a grid between 1.00 and 1.015 at intervals of 0.001. Specific gravity is defined as a ratio of the density of a solution to the density of water. Within the metabolite data, this means that each sample will have a specific gravity that equals or exceeds 1. We observed similar results for alternative sample quality measures based on creatinine concentrations and the number of metabolites below their fifth percentile per sample (see Supporting Information).

2.5 | Predictive modeling methods

2.5.1 | Initial variable selection and metabolite dichotomization

Some metabolites in our data set had missing abundance values because their abundances fell below the LOD. Since imputed values do not accurately reflect the true metabolite concentrations, they can negatively affect the estimation of the predictive model. To address this problem, we dichotomized the values of those metabolites with a high number of observations below their respective LODs. For every metabolite, we first built a contingency table between the dichotomized metabolite and the binary TCMR class response. We then performed a Fisher exact test to assess if there was a significant association between the TCMR and the "absence" of the metabolite. Since we performed multiple Fisher exact tests (one per metabolite), we adjusted our *P*-values into *q*-values using the Benjamini-Hochberg FDR correction method and set the *q*-value threshold to limit the expected number of false discoveries to 4. If a significant association exists and more than 25% of the metabolite values fall below its LOD, then we dichotomized the metabolite prior to modeling. If a significant association does not exist and more than 25% of the metabolite values fall below its LOD, we removed the metabolite. Otherwise, if neither set of conditions is satisfied, we used the metabolite concentration as a continuous variable.

2.5.2 | Random up-sampling of the positive class in the TCMR data

Due to the class imbalance present in the TCMR data, where there are many more non-TCMR samples than TCMR samples, we implemented random up-sampling of the TCMR samples within our training sets (see Section 2.6 on the cross-validation approach used). The intention of this approach was to allow the models to be built upon a less imbalanced training set, and thus potentially improve the overall predictive performance on unseen data. For each training set, we randomly sampled (with replacement) exactly 45 TCMR observations found within the training set and included those duplicated observations within the training set. By doing this for 45 observations, we approximately doubled the number of TCMR observations present within each training set. We avoided attempting to up-sample even more TCMR observations (to achieve a closer balance between both classes) due to the risks of model overfitting and exacerbating the class imbalance severity.

2.5.3 | Estimation methods

We used eight different estimation methods to build our predictive TCMR model. These estimation methods are as follows: (1) least absolute shrinkage and selection operator (LASSO), (2) ridge, (3) elastic net (EN), (4) post-elastic net (PostEN), (5) partial least squares (PLS), (6) random forest (RF), (7) support-vector machine (SVM), and (8) XGBoost (XGB). LASSO,[35] ridge,[36] and elastic net are all regularized regression methods that constrain coefficient sizes.[37] LASSO regularizes the sum of the absolute value of the coefficients, which permits variable selection by reducing the size of some coefficients to zero. Ridge regularizes the sum of the squared coefficients, and is useful to estimate models that contain correlated

covariates (also known as multicollinearity) which is observed in metabolite data. Elastic net combines both types of regularization, achieving the properties of both. While these regularization methods can improve model performance (compared to classical regression methods), they introduce bias in the estimated coefficients. Post-elastic net uses elastic net to select variables and then fits either a ridge regression (for multiple selected variables) or a classical logistic regression (for a single selected variable) on the selected variables.

Partial least squares regression is similar to principal component regression,[38] and both are useful techniques to handle high dimensional data and potential multicollinearity. The latter method decomposes the explanatory data to a set of principal components, while the former decomposes both the explanatory data and the response data (maximizing the covariance between their decompositions). PLS is quite commonly used in metabolomics.[39] In this article, we specifically use partial least squares-discriminant analysis (PLS-DA) since the response is a binary variable, but hereon use the shorter PLS abbreviation.

Random forest is a machine learning technique that builds many decision trees.[40] To build each decision tree, a random subset of predictors is taken at the initial node. Among this collection of predictors, the predictor that best classifies the samples is chosen to divide the data into multiple branches that terminate into nodes. We repeat this process until each of the terminal nodes contain a single sample. To avoid over-fitting the data, each decision tree is trained using a random subset of the data. When predicting the class of a sample using a random forest, each decision tree votes on a class and the class probabilities become the proportion of trees voting for that class.

Support-vector machine[41,42] and XGBoost[43]are the final two nonlinear machine learning methods that we use. Support-vector machine aims to build a hyperplane that effectively divides the observations of two classes, such that the margin of separation between these two classes is maximized. This method has been further extended to allow for nonlinear classification by Boser et al,[41] through the use of nonlinear kernels. We specifically use a quadratic polynomial kernel when constructing the prediction model with this estimation method. XGBoost is a relatively novel method that creates an ensemble of decision trees (similar to random forest). However, it does so through the use of a gradient boosting framework (to create gradient boosted trees) and was developed to allow for numerous modeling benefits such as scalability to very large datasets, handling of missing data, and other computational features.

2.6 | Assessing prediction performance

In the absence of an independent test set, we used cross-validation to assess the prediction performance of the models so that we could generalize our results to similar types of datasets. An additional benefit of using cross-validation to assess prediction performance is that we do not need to withhold a substantial fraction of the data for a test set. Since our data contains multiple samples from each patient, we used a "patient-based cross-validation" protocol that generates the folds from the 59 patients, rather than the total number of samples (334). Using a sample-based cross-validation approach may cause samples from the same patient to appear in both the training set and the validation set which can result in an overestimation of the model's prediction performance. We specifically used leave-one-out patient-based cross-validation (LOOCV), such that each fold contains a single patient's set of samples. For each training set during LOOCV, we used the methods described in Section 2.5 (including standardization of the predictors) to train the predictive model and obtain predictions from the corresponding validation set. To set the particular hyperparameters for each of the estimation methods, we used a traditional 5-fold cross-validation protocol nested within the LOOCV (nested cross-validation).

After completing LOOCV, we obtained class predictions for each of the 334 samples to generate the receiver operating characteristic (ROC) curve and to compute the area under the curve (AUC). Lastly, we repeated the above outlined procedure for 100 repetitions to account for variability in model performance estimates, attributed to different fold splits from the nested 5-fold cross-validation. Due to the computational complexity, only one repetition was performed when examining the numerous different sample quality thresholds.

3 | RESULTS

3.1 | TCMR data: Demographic and clinical characteristics of patients

Samples from patients in the TCMR study were classified using the classification scheme described in Section 2.2, resulting in 45 "TCMR" samples and 289 "non-TCMR" samples. Of the 59 patients in the study, 19 fell under the "patients with

TABLE 1 Medical history continuous variable summary statistics

	Age at transplant	Age of donor	Years on dialysis
Mean	11.44	33.24	1.36
SD	4.68	11.96	1.39

Note: Summary statistics on some of the medical history variables for the 59 patients in this study. The units for each variable are in years.

TABLE 2 Principal components

	PC1	PC2	PC3	PC4	PC5
SD	5.94	5.63	3.08	2.74	2.47
Proportion of variance	27.12%	24.40%	7.32%	5.76%	4.70%
Cumulative proportion	27.12%	51.52%	58.83%	64.59%	69.30%

Note: SD of the first five principal components (PC) and their proportions of the total variance of the metabolite dataset.

at least one TCMR sample" category while the remaining 40 fell under the "Patients with non-TCMR samples" category. Among the 59 patients, 25 are female and 34 are male. Also, the majority of patients are either Caucasian (n = 32) or Canadian First Nations ethnicity (n = 19), while the remaining patients come from a mixture of different racial groups. The two most common original diseases that the patients had, which led to a kidney transplant, were dysplasia/hypoplasia (n = 17) and juvenile nephronophthisis (n = 6), with the remaining original diseases leading to a kidney transplant being relatively uncommon (n = 3 or fewer patients per disease) or unknown (n = 3). For each of the race and original disease variables, we combine categories smaller than the two largest categories into a single "Other" category because of their much lower frequencies. The donors for the kidney transplants were either deceased donors (n = 28) or living donors (n = 31). The remaining summary statistics for the continuous variables can be found in Table 1.

3.2 | TCMR metabolomic data: Exploratory analysis

Prior to building different classifiers to predict TCMR from urinary metabolomic data, we generated a heat map of the metabolite values to explore the overall quality and structure of the data (see Figure S1 in the Supporting Information). Not surprisingly, we observe some clear clustering of metabolites indicative of the expected high correlations among certain metabolite concentrations.

In addition, we explored the metabolomics data by first performing a principal component analysis (PCA) to reduce the data dimensionality while maintaining most of its total variation. Although this preliminary analysis is not necessarily informative of prediction performance, PCA is a powerful method to explore potential problems in the data (eg, batch effects) and to discover groups among samples. In this analysis, we matched the 19 patients with at least one TCMR sample with one of the 40 patients without any TCMR event. We also took a single sample from each patient, resulting in a total of 38 samples being available for PCA. Prior to standardization, two metabolites were removed because all 38 of their respective values were the same value (preventing standardization).

The results of the PCA (for the first five principal components) can be observed in Table 2 and Figure S2. The first five principal components represent nearly 70% of the total variability in the metabolite dataset. This is likely reflective of the high correlations between many of the different metabolites. The first two principal components barely account for a majority of the variability (slightly over 50% of the cumulative variance) among the 130 metabolites in these 38 samples. Thus, the different groups cannot be clearly identified in a scatter plot of these components (Figure S2). Outliers of the first category tend toward larger negative values of the second principal component while the majority of outliers of the second category tend in the opposite direction.

3.3 | Effect of normalization and logarithmic transformation on prediction of TCMR

Although many normalization techniques have been proposed and studied in metabolomics, normalization to creatinine is a simple scaling technique used in many urinary metabolomic studies to account for a potential dilution effect in urinary

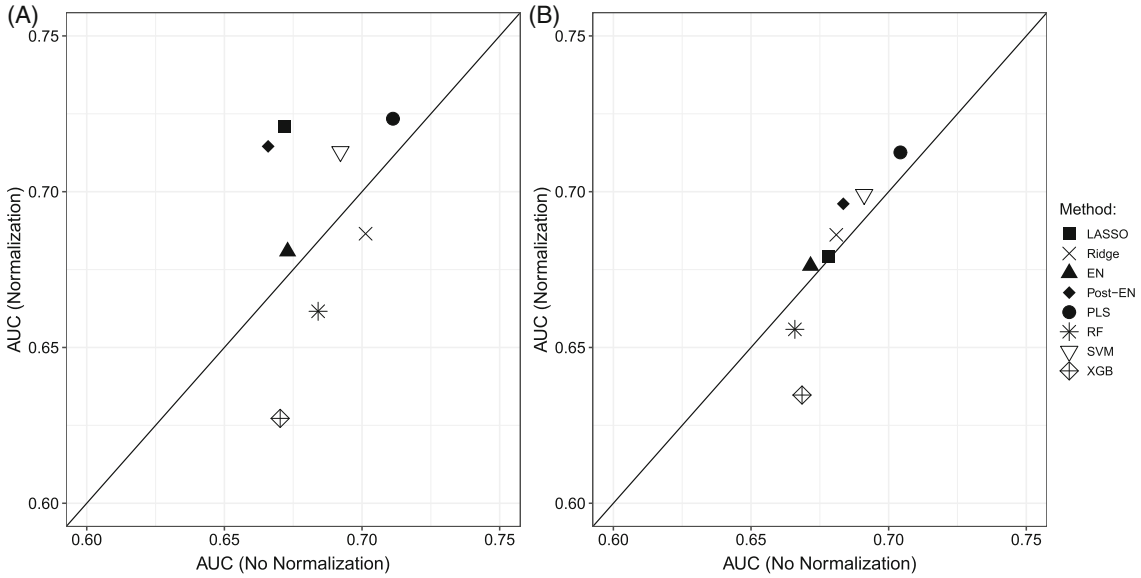

FIGURE 1 Creatinine normalization vs no normalization scatterplots. Scatterplots of the mean AUCs of the estimation methods based on non-normalized and normalized TCMR metabolite data. The x-axis and y-axis represent the AUCs resulting from an LOOCV using non-normalized and normalized metabolite data, respectively. A 45° line has been drawn as a reference. If a point lies above the line, then that indicates the model performs better when the metabolites are normalized by creatinine. If a point lies below the line, then that indicates the model performs better without creatinine normalization. (A) Corresponds to the results of modeling on nontransformed data while (B) corresponds to the results of modeling on log-transformed data

samples. Despite the simplicity of using a ratio to normalize each metabolite, the subsequent statistical analyses of the normalized data can become more complex and obscure. Another common processing step is the use of a logarithmic transformation to reduce the heteroscedasticity or skewness usually observed in urinary metabolite concentrations. Since these normalization and transformation techniques are widely used in clinical and biomedical studies, in this section, we explore in detail their effects on the performance of various classification methods using the TCMR metabolomic data.

Figure 1 shows the mean AUC (across 100 LOOCV runs) for the eight classifiers we built to compare their prediction performances across different pre-processing settings. AUC values typically range from 0.5 (indicating a very poor predictive model, see Figure S3) to 1 (indicating a perfect predictive model). Not surprisingly, some methods perform better than others, regardless of the preprocessing setting. For example, PLS has the best performance in both non-normalized and normalized metabolomic data, before and after a logarithmic transformation. Interestingly, LASSO and PostEN, which perform a variable selection as coefficients are estimated, seem to be more sensitive to the effect of normalization on the raw data. In both cases, the prediction performance improves when metabolite ratios (concentrations relative to creatinine) are used to build and test the classifiers. For other methods, the creatinine normalization does not appear to have a remarkable impact on performance. In general, there are marginal differences in prediction performance between models trained on normalized data and models trained on non-normalized data after a logarithmic transformation of the data.

As explained and illustrated by Curran-Everett, a ratio will not normalize the numerator if the numerator is not strictly proportional to the denominator.[23] If the ratio fails to normalize some metabolite concentrations to the creatinine concentration, using metabolite ratios can adversely affect the results and create group differences that do not exist or mask existing ones that do exist. The concentrations of some metabolites in the TCMR dataset are not proportional to creatinine levels and thus are not suitable for this kind of normalization. For example, Figure 2A shows that the concentration of PC.ae.C44.4 is not related to the concentration of creatinine, with many low values observed at different creatinine levels. The relative concentration for this metabolite conveys a different message than that present in the original data, which may distort a class prediction when used as a covariate.

On the other hand, for the informative metabolites in the TCMR dataset that are proportional to creatinine levels, this normalization appears to be effective and results in an improvement of prediction performance. The effects of normalization seem to be more pronounced in those methods that build a classifier using only a subset of selected features (eg, LASSO) as we observed in our data (Figure 1A). Figure 2B shows the frequency of selection for the 20 most frequently selected metabolites as determined by LASSO in 5900 training sets based on both the non-normalized and the normalized

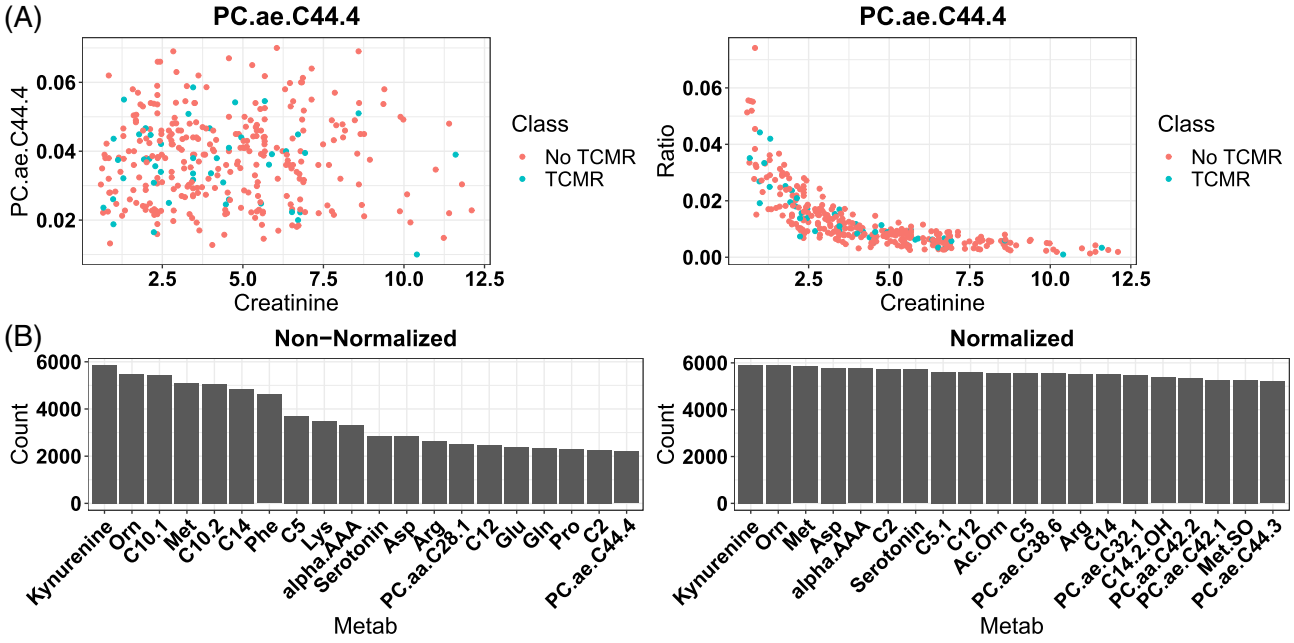

FIGURE 2 Scatterplots of PC.ae.C44.4 and Barplots of LASSO Variable Selection Frequencies. Panel A presents two scatterplots of the metabolite PC.ae.C44.4 against creatinine, the first when non-normalized and the second when normalized. Panel B shows the results of LASSO variable selection frequencies from 5900 training sets (59 folds * 100 repetitions of the LOOCV), for non-normalized and normalized TCMR metabolite data. The top 20 metabolites are displayed in each plot

data (ie, 100 runs of a 59-fold LOOCV). While the impact of normalization may vary greatly among metabolites in this dataset, overall, we observe a more stable selection when the data is normalized. This results in an improved overall performance for this method (Figure 1A). Methods that do no select variables, such as PLS and SVM seem to be less affected by the normalization protocols (Figure 1A).

In general, the effects of the normalization seem to be reduced by the logarithmic transformation (Figure 1B), with most classifiers having most the same performance in normalized and non-normalized logged data. Nevertheless, it is important to note that, in some cases, we are introducing a remedy with a scaling technique that has failed to normalize the concentration of many metabolites to begin with.

Since this naive scaling method does not appear to have a consistent effect on metabolites and more effective methodologies have been proposed to normalize metabolomic data,[20,21] we do not recommend to use it blindly in discovery studies. While the logarithmic transformation can also be problematic (especially in low abundant metabolites), in general, it helps to control the effect of outliers and skewness in the overall prediction performance of most classifiers. In Section 3.6 we explore the generalizability of these observations in other independent datasets.

3.4 | Effect of sample quality thresholds on TCMR prediction performance

Based on the results described in the previous sections, here, we examined the effect of eliminating low quality samples from the non-normalized logarithmic metabolomic dataset. Considering the computational complexity of building so many classifiers we decided to focus this analysis on only a few modeling techniques. In particular, we selected one variable selection method (PostEN), one linear method without variable selection (PLS), and one nonlinear method (SVM) that showed competitive performances in all settings (Figure 1). Given their relatively poor performance in our dataset, we did not include the tree-based classifiers (ie, RF and XGBoost) in this analysis.

Using specific gravity as a sample quality metric, a scatterplot was initially generated to demonstrate how the number of "TCMR" and "Non-TCMR" samples changed as the quality control threshold increases (Figure 3A). Figure 3B illustrates how the performance of the classifiers changes as the most diluted samples are excluded from the analysis. Across all methods, it appears that the performance mostly plateaus (absence of a relevant peaking) before decreasing drastically due to a small number of observations to train the model. For this dataset, we do observe a minor increase in prediction

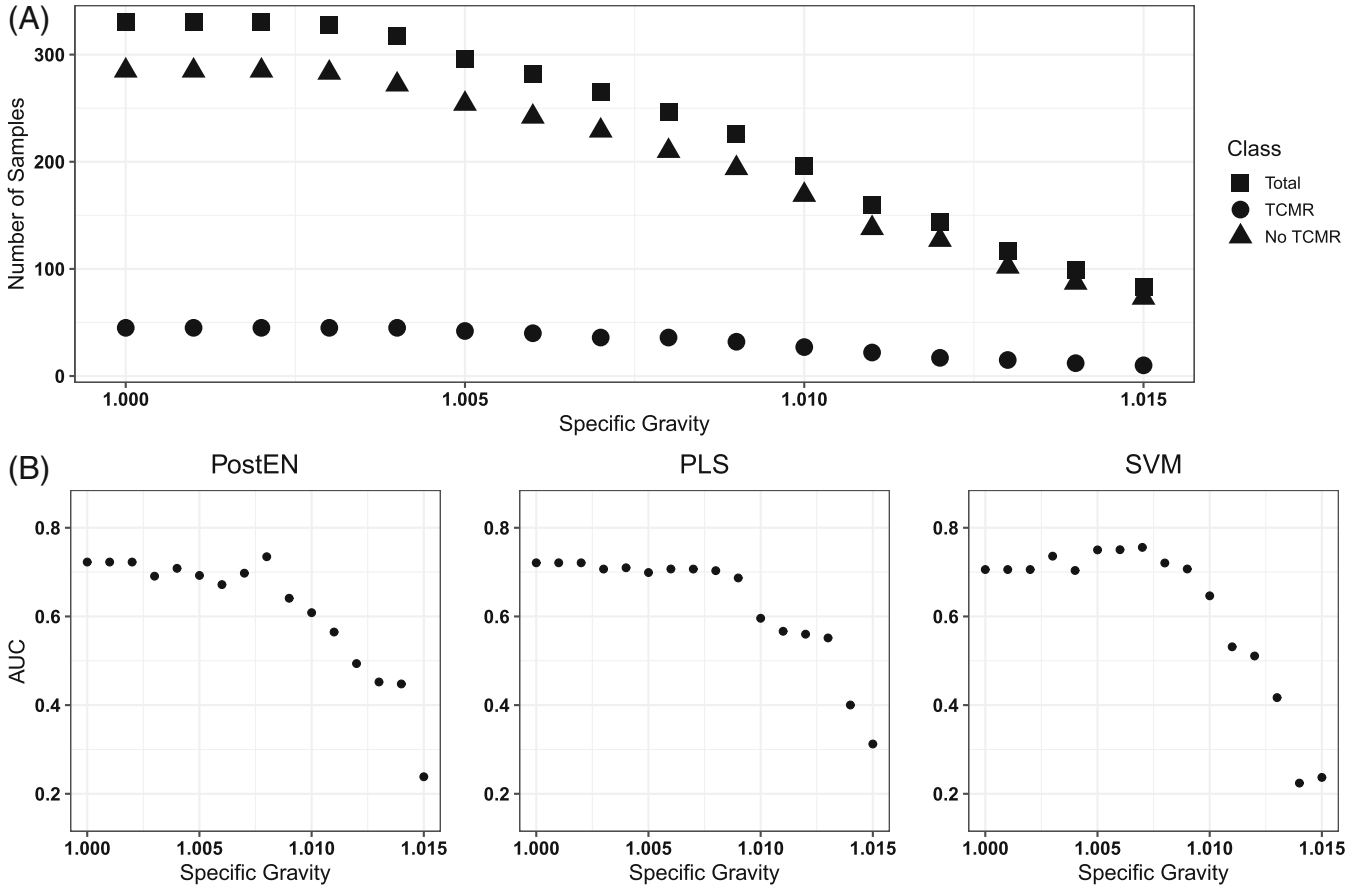

FIGURE 3 Specific gravity threshold analysis. (A) Shows the number of samples as a function of specific gravity threshold (ie, samples that have specific gravities that exceed the threshold). The legend indicates the different labels for the total number of samples, the number of "TCMR" samples, and the number of "No TCMR" samples. (B) Displays the results of the specific gravity threshold analysis for PostEN, PLS and SVM on non-normalized logarithmic metabolite data, with points corresponding to AUCs from a single LOOCV repetition

performance for the SVM model if only those samples with a specific gravity above 1.005 are used in the analysis. However, we believe that the gain in performance is marginal and comes at the cost of losing about 34 samples (three of which are "TCMR" samples). These results are consistent with those obtained for other methods and using other quality measures (Figures S4-S12).

3.5 | TCMR classifiers: With and without medical history data

In this section, we analyzed some TCMR classifiers in more detail and continue to focus on PostEN, PLS and SVM as representative methodologies in the different classes explored. In particular, we examined the concordance among the most important features of the classifiers built from non-normalized logarithmic metabolomic data, and we explored how to complement these classifiers with medical history data.

As previously observed, some classifiers have very competitive performances predicting TCMR classes. Regardless of their similar performances, it is interesting to compare the concordance among the most relevant variables used by each method to predict TCMR classes. Among all the methods that we tested, only LASSO and EN select a strict subset of variables to build the classifier. Thus, instead of comparing the set of selected variables, we created a (mean) ranking across the 5900 classifiers built for each method based on a variable importance measure.[44] For LASSO and PostEN, the variable importance was given by the absolute sizes of the coefficients. For PLS we used the Variable Importance in the Projection (VIP). Since we used a nonlinear kernel for SVM, without recursive feature elimination, we did not include this method in the comparison. The resulting mean rankings were used to compute the area under the concordance curve

TABLE 3 Classifier AUCs

Estimation method	Metabolite classifier	Composite classifier	Ensemble classifier
PostEN	0.684	0.689	0.666
PLS	0.704	0.706	0.690
SVM	0.691	0.688	0.682

Note: The mean AUC for each estimation method and classifier combination, across the 100 repetitions of patient-based cross-validation. The results are based on non-normalized and logged TCMR metabolite data and TCMR medical history data (the latter used in the composite and ensemble classifiers).

(AUCC) between methods.[45,46] The area under the concordance curve (AUCC) is 0.61 between LASSO and PLS, 0.86 between LASSO and PostEN, and 0.54 between PLS and PostEN.

A natural way to complement the metabolomics data is to include the medical history data of the patients. Thus, we used medical data in conjunction with the non-normalized logged metabolite data to create two additional TCMR classifiers. For each method, we built and tested four classifiers: a pure metabolite classifier (which uses only the metabolite data), a pure medical history classifier (which uses only the medical history data), a composite classifier (which uses both the metabolite and medical history data together), and an ensemble classifier (which averages predictions from the metabolite and medical history classifiers). Table 3 compares the prediction performances of the resulting classifiers. The performance results for classifiers built using only medical history data were omitted from the table because of their generally poor prediction performance, with AUCs never exceeding 0.55. For both PostEN and PLS methods, the mean AUCs improve only very slightly for the composite classifier and are worse for the ensemble classifier. However, the metabolite classifier built using SVM has a better mean AUC than both the composite and ensemble classifiers, with the ensemble once again giving the lowest mean AUC.

3.6 | Generalizability of the effect of preprocessing techniques

In Section 3.3, we observed that rescaling metabolite concentrations with creatinine levels in our TCMR data may not properly scale all metabolite concentrations as intended and may have unforeseen effects on the prediction performance of a classifier. In addition, a logarithmic transformation can be a remedy to the problems introduced by the normalization, as well as potential problems of skewness and outliers in some metabolite distributions. In this section, we examine the effect of these preprocessing techniques using three additional urinary datasets collected for different medical conditions to assess the generalizability of our observations.

The three additional datasets that we used are: (1) the IC/PBS dataset, (2) the MMF dataset and (3) the DMD dataset. We did not have access to any potentially identifying personal information.

1. IC/PBS is a urinary LC-MS metabolite intensity dataset, collected with the aim of diagnosing patients with interstitial cystitis/painful bladder syndrome (IC/PBS). It contains a total of 92 samples (56 cases and 36 matched controls) and complete data (ie, no missing values) for 490 metabolites measured with LC-MS.
2. MMF is a urinary LC-MS MMF is a quantitative urinary LC-MS metabolite dataset, collected in pediatric transplant recipients at the time of pharmacokinetic testing, to identify a metabolomic predictor of mycophenolate mofetil (MMF) pharmacodynamics relative to measured drug exposure. It contains a total of 52 samples (33 cases and 19 controls) and a total of 133 metabolites, with missing values imputed by each metabolite's limit of detection divided by the square root of 2 (ie, LOD/sqrt (2)).
3. DMD is a urinary LC-MS metabolite data that was collected to classify patients with Duchenne muscular dystrophy (DMD). It contains a total of 51 samples (39 cases and 12 controls) and a total of 24 metabolites, with missing values imputed by half of each metabolite's minimum value.

Figures 4, 5 and 6 illustrate the ROC curves (with median AUCs) for PostEN, PLS, and SVM, respectively, for the four LC-MS urinary metabolomic datasets preprocessed with different combinations of techniques (ie, with and without creatinine normalization, and with or without logarithmic transformation). Overall, the normalization appears to increase the performance of the three classifiers in TCMR and MMF datasets. In particular, the improvement is more relevant for PostEN on logged metabolite data, with the median AUC increasing from 0.681 to 0.747 in TCMR and 0.636 to 0.654

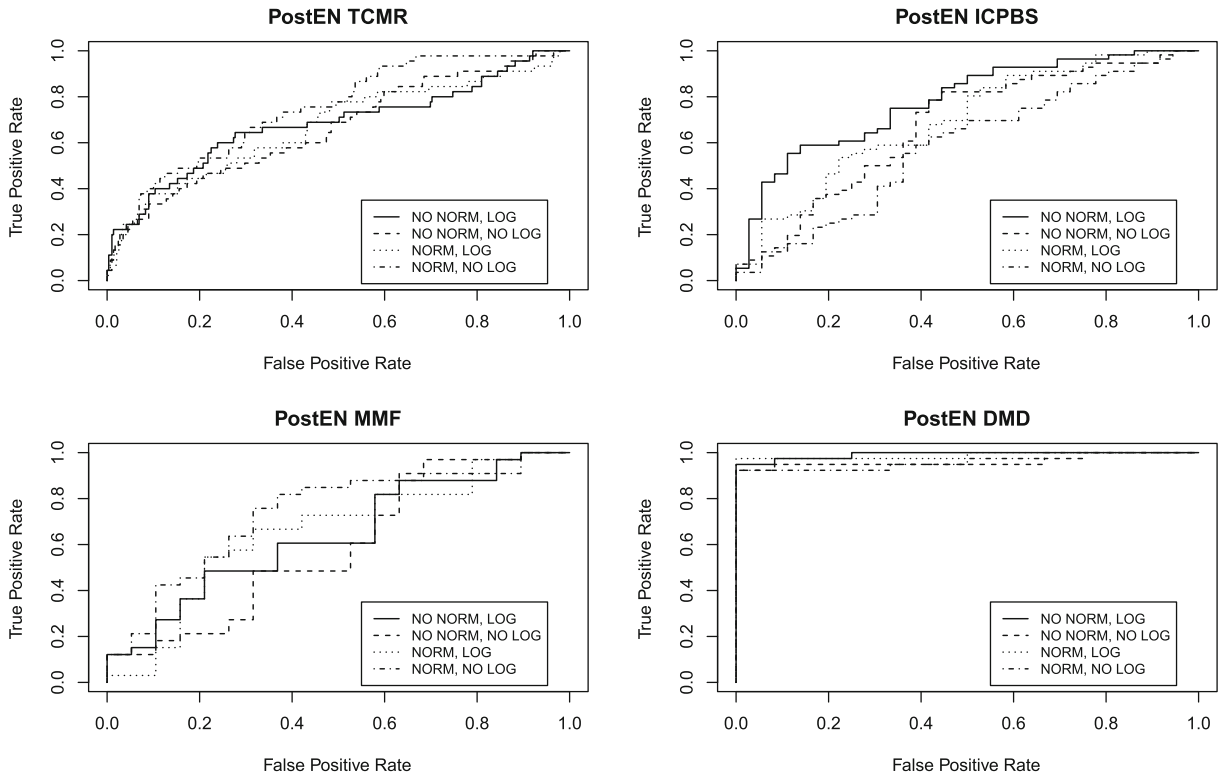

FIGURE 4 Effect of preprocessing techniques in four different datasets (PostEN). ROC curves of the PostEN model for different combinations of normalization and log transformations, with each plot corresponding to one of the four datasets examined. Each ROC curve corresponds to the curve that achieves the median AUC among 100 repetitions of LOOCV

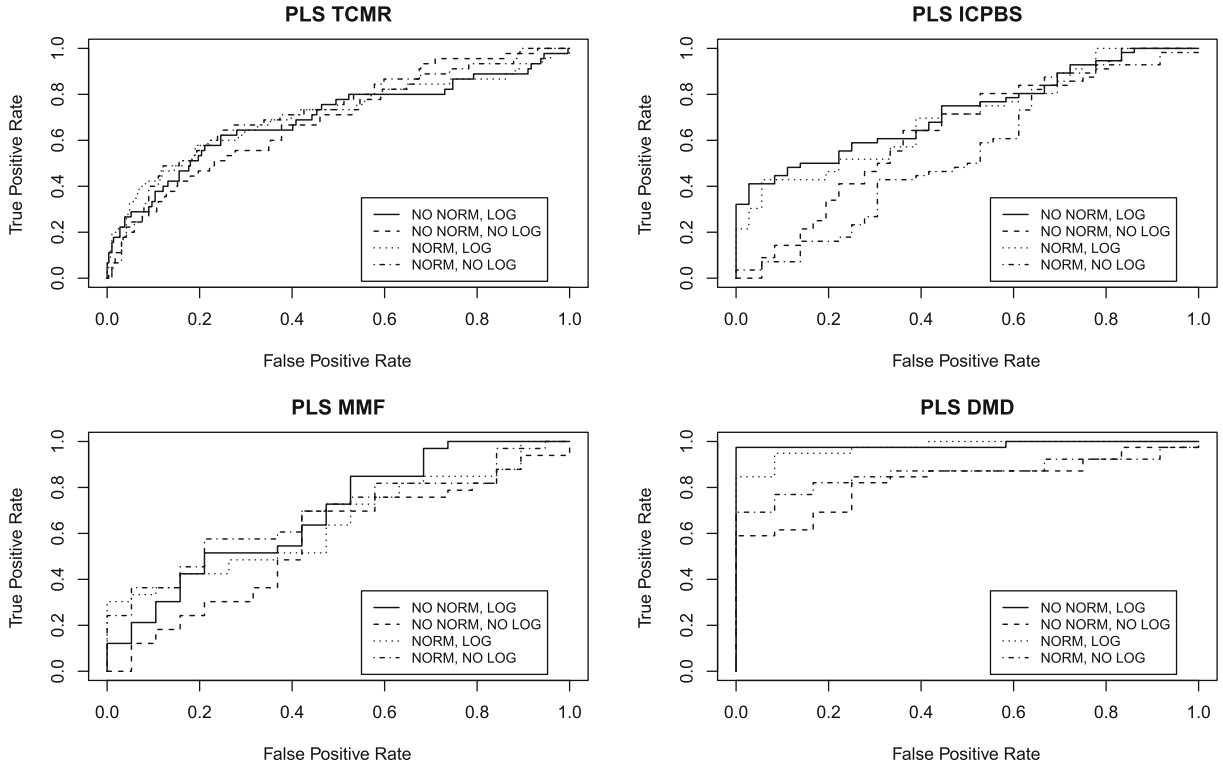

FIGURE 5 Effect of preprocessing techniques in four different datasets (PLS). ROC curves of the PLS model for different combinations of normalization and log transformations, with each plot corresponding to one of the four datasets examined. Each ROC curve corresponds to the curve that achieves the median AUC among 100 repetitions of LOOCV

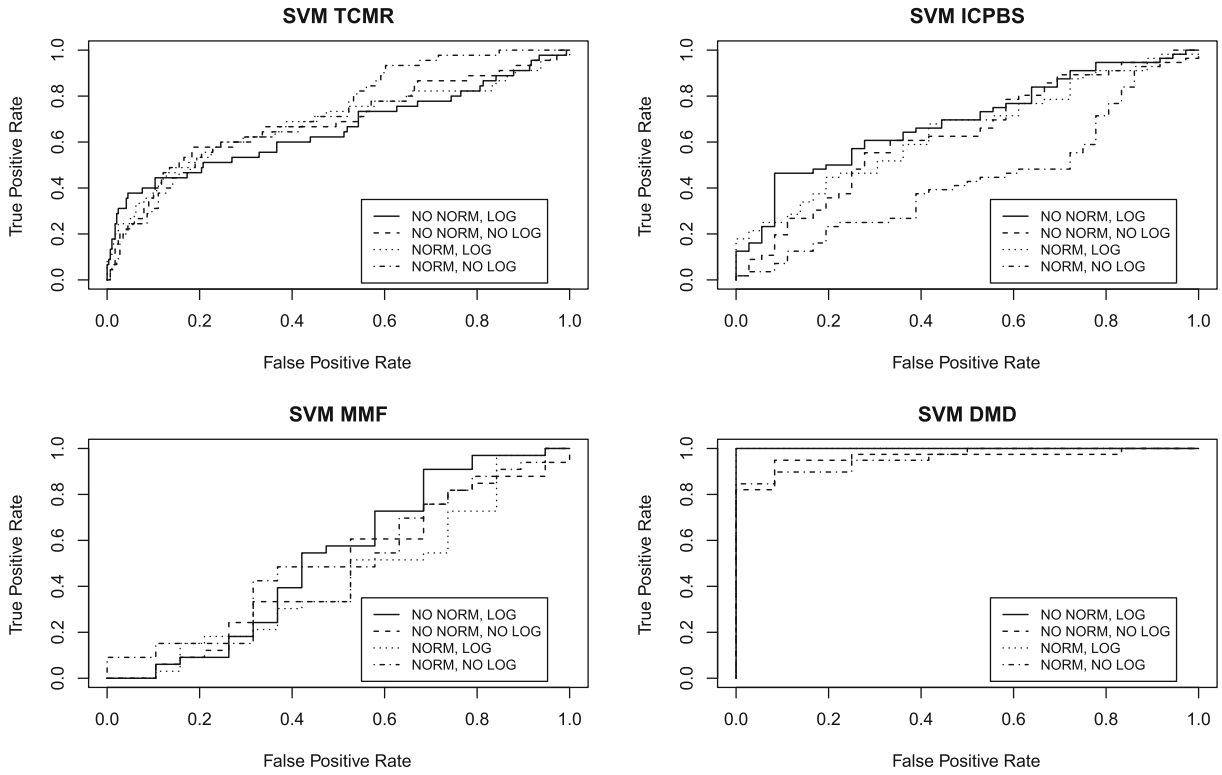

FIGURE 6 Effect of preprocessing techniques in four different datasets (SVM). ROC curves of the SVM model for different combinations of normalization and log transformations, with each plot corresponding to one of the four datasets examined. Each ROC curve corresponds to the curve that achieves the median AUC among 100 repetitions of LOOCV

in MMF. As previously observed in TCMR data, the normalization affects in particular those methods that select a subset of metabolites to classify. The stabilization in LASSO variable selection observed for TCMR is still present for MMF (Figure S15B), which may be attributed to the presence of many metabolites proportionally related to creatinine (eg, Dopamine in Figure S15A). PLS and SVM, which use all metabolites (including creatinine) to build a classifier, are less sensitive to the effect of the normalization in both datasets.

Interestingly, we observe different effects after normalizing the IC/PBS and the DMD data. The classification performance of the three methods is negatively affected by the normalization, especially for IC/PBS. Unlike in previous datasets, the selection of variables by LASSO becomes more unstable after normalization (see Figure S16B). It is important to note that these datasets are different in nature. The IC/PBS contains very large intensity values, which are not proportional to creatinine in most cases (see Figure S16A). Thus, in general, creatinine cannot be used to properly scale the metabolite concentrations. In this case, the normalization negatively affects most multivariate analyses. For the DMD study, creatinine is part of the disease process and a main marker of the loss of muscle mass in DMD, violating an important assumption of creatinine normalization (see Figure S17). Nevertheless, this dataset contains only some predefined targeted metabolites measured with high level of precision and accuracy, which explains the high AUC values in all cases. Results still need to be validated in a larger cohort which potentially examines the effects of steroid use, dystrophin mutation, and/or loss of ambulation.

As mentioned before the logarithmic transformation tends to counter act potential effects of the normalization. While the effects of the logarithmic transformation also depend on the distribution of the data being transformed, in general, it had a positive effect in the performance of most classifiers and datasets.

4 | DISCUSSION

Developing predictive models of TCMR in renal transplant patients remains a challenging problem, even with the availability of urinary metabolite data. However, studying different methodologic approaches may reveal data processing

techniques that can help improve the performance of current TCMR predictive models. In this study, we explored some common preprocessing techniques in combination with different machine learning tools to build classifiers and assess their impact on prediction performance.

Despite the simplicity and widespread use of creatinine normalization, this technique may not always be appropriate to scale the signals of many metabolites in urinary metabolomic data. Overall, we observed that this normalization technique has inconsistent and unforeseen effects and their impact on classification performance differ based on the methods and the data. Some classification methods, like LASSO, may be highly impacted by its (positive or negative) effects since their classification is based on only a subset of selected metabolites. In general, using metabolite ratios (relative to creatinine) should not be used without previously studying the distribution of the metabolite measurements and their relation to creatinine. Other normalization techniques may be more appropriate depending on the dataset to be analyzed.[19-21] Given the skewness and presence of outliers usually present in metabolomic data, a logarithmic transformation can help to reduce some data variability and generate a more Gaussian distribution. While the logarithmic transformation can be problematic in the presence of many low abundant metabolite values, we observed an overall improvement in classification performance for most methods and most datasets analyzed.

Other preprocessing methods that merit further investigation in future work are alternative ways to better address the left-censorship in metabolite levels. In particular, methods that distinguish between different types of zeros or missing values before imputation can be useful to guide the analysis of metabolomics data.[47-49] However, their implementation should be tailored to the characteristics of the metabolomics data being analyzed, including units of measurements, availability of raw measurements below limit of detection, estimates of limit of detection of different metabolites, and availability of additional metadata, among others.

The sample quality analysis in the TCMR dataset shows that none of the sample quality metrics used to filter potential low-quality samples had a substantial effect on the prediction performance of the models. In fact, removal of "low quality" samples may result in information loss since the model trains on fewer samples and is only generalizable to "high quality" samples in new data. Other sample quality metrics could be investigated, including osmolality. While specific gravity is often used as an estimate of osmolality, it may over- or underestimate this metric.[50] The inclusion of demographic data was not very informative to classify TCMR samples. We observed some minor improvements in the composite classifiers compared to the metabolite and the ensemble classifiers. Overall, the main goal of this analysis was to demonstrate different ways of integrating the data from different sources. Results will largely depend on the nature and richness of the information being integrated.

As expected, the relative performance of the different classifiers varies across datasets. In general, PostEN and PLS had the best classification performances in all non-normalized logged datasets. The primary disadvantage of PLS is the lack of variable selection and its sensitivity to multicollinearity problems in most metabolomic datasets.[51] PostEN inherently performs variable selection and accounts for multicollinearity problems, resulting in a high performance for all the datasets analyzed in this study. Overall, we observed moderate concordance of the features selected or highly weighted by these methods.

The tree-based classifiers RF and XGBoost had poor performances in some datasets due, probably, to the presence of many uninformative and unusually noisy metabolites. In high-dimensional datasets, some feature subsets sampled at specific decision tree nodes may not contain any variables that effectively predict the response, resulting in poor node splits. Increasing the proportion of sampled variables to alleviate this issue can instead result in high correlations among decision trees, which can also impair performance for methods like RF.[40] There have been some proposed methods that have attempted to resolve this issue for high-dimensional datasets, via weighted sampling of the predictors.[52,53] Additionally, the low performances achieved by XGBoost could have also stemmed from the extensive tuning required for many of its hyperparameters.[43,54] In our analyses, we performed some hyperparameter tuning when applying the XGBoost method, but its computational demands lead us to restrict our analyses to only search through small grids of reasonable hyperparameter values. More focused research into the application of XGBoost in the metabolomics context could provide additional useful insight.

While our paper is mainly focused on the analysis of metabolomic data to enhance the prediction of TCMR, we investigated the effect of common preprocessing techniques in three additional datasets to illustrate the complexity of the problem and the diversity of results. It is important to note that the goal of this study was not to evaluate the clinical utility of the classifiers or selected markers. Additional validation studies would be required to assess the benefits of the identified metabolomic panels and to interpret their biological relevance. Overall, we believe the statistical and computational techniques reviewed in this study are applicable to a wide range of -omics studies in health sciences. Rigorously studying all the analytical steps of complex biomarkers studies is essential for clinicians to have confidence in those models

that will ultimately be used to make critical decisions in the clinic. Increasingly, integrated use of multiple biomarkers together will improve diagnostic precision and personalize care, if they can be properly processed and suitably modeled.

ACKNOWLEDGEMENTS

Most of the statistical analysis was done using GCF's computational infrastructure funded by the Canada Foundation for Innovation (CFI). Kevin Multani was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Undergraduate Student Research Award (USRA) grant and GCF's NSERC Discovery grant. Nikolas Krstic was supported by GCF's NSERC Discovery grant. The TCMR analysis was supported by an operating grant from CHIR (#274755) and the MMF pharmacometabolomic analysis was supported from a Biomedical Research Grant from Kidney Foundation of Canada (KFOC170003). The DMD dataset was generated as part of the NIH study (AR056973) and LC-MS was performed at the Southeast Center for Integrated Metabolomics (U24DK097209).

DATA AVAILABILITY STATEMENT

TCMR, MMF and DMD data are available on request due to privacy/ethical restrictions. The IC/PBS data is available at the NIH Common Fund's National Metabolomics Data Repository (NMDR) website, the Metabolomics Workbench, https://www.metabolomicsworkbench.org, where it has been assigned Project ID PR000299. The data can be accessed directly via its Project DOI: 10.21228/M8D60P. This work is supported by NIH grant, U2C- DK119886.

ORCID


Nikolas Krstic https://orcid.org/0000-0002-8080-3756

REFERENCES

1. Burke HB. Predicting clinical outcomes using molecular biomarkers. *Biomark Cancer*. 2016;8:89-99. doi:10.4137/BIC.S33380
2. Chen JJ, Lu TP, Chen YC, Lin WJ. Predictive biomarkers for treatment selection: statistical considerations. *Biomark Med*. 2015;9(11):1121-1135. doi:10.2217/bmm.15.84
3. Weiner J 3rd, Maertzdorf J, Sutherland JS, et al. Metabolite changes in blood predict the onset of tuberculosis. *Nat Commun*. 2018;9(1):5208. doi:10.1038/s41467-018-07635-7
4. Kauppi AM, Edin A, Ziegler I, et al. Metabolites in blood for prediction of bacteremic sepsis in the emergency room. *PLoS One*. 2016;11(1):e0147670. doi:10.1371/journal.pone.0147670
5. Huang S, Chong N, Lewis NE, Jia W, Xie G, Garmire LX. Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. *Genome Med*. 2016;8(1):34. doi:10.1186/s13073-016-0289-9
6. Zhao Y, Lv H, Qiu S, Gao L, Ai H. Plasma metabolic profiling and novel metabolite biomarkers for diagnosing prostate cancer. *RSC Adv*. 2017;7(48):30060-30069. doi:10.1039/c7ra04337f
7. Liu X, Zhang M, Liu X, et al. Urine metabolomics for renal cell carcinoma (RCC) prediction: tryptophan metabolism as an important pathway in RCC. *Front Oncol*. 2019;9:663. doi:10.3389/fonc.2019.00663
8. Luck M, Bertho G, Bateson M, et al. Rule-mining for the early prediction of chronic kidney disease based on metabolomics and multi-source data. *PLoS One*. 2016;11(11):e0166905. doi:10.1371/journal.pone.0166905
9. Rush D, Somorjai R, Deslauriers R, Shaw A, Jeffery J, Nickerson P. Subclinical rejection--A potential surrogate marker for chronic rejection--may be diagnosed by protocol biopsy or urine spectroscopy. *Ann Transplant*. 2000;5(2):44-49.
10. Wang JN, Zhou Y, Zhu TY, Wang X, Guo YL. Prediction of acute cellular renal allograft rejection by urinary metabolomics using MALDI-FTMS. *J Proteome Res*. 2008;7(8):3597-3601. doi:10.1021/pr800092f
11. Blydt-Hansen TD, Sharma A, Gibson IW, Mandal R, Wishart DS. Urinary metabolomics for noninvasive detection of borderline and acute T cell-mediated rejection in children after kidney transplantation. *Am J Transplant*. 2014;14(10):2339-2349. doi:10.1111/ajt.12837
12. Janeway CA Jr, Travers P, Walport M, Shlomchik MJ. Immunobiology: The Immune System in Health and Disease. *The major histocompatibility complex and its functions*. 5th ed. New York, NY: Garland Science; 2001. Available from: https://www.ncbi.nlm.nih.gov/books/NBK27156/
13. Ingulli E. Mechanism of cellular rejection in transplantation. *Pediatr Nephrol*. 2010;25(1):61-74. doi:10.1007/s00467-008-1020-x
14. Halloran PF, Reeve JP, Pereira AB, Hidalgo LG, Famulski KS. Antibody-mediated rejection, T cell-mediated rejection, and the injury-repair response: new insights from the genome Canada studies of kidney transplant biopsies. *Kidney Int*. 2014;85(2):258-264. doi:10.1038/ki.2013.300
15. Marcussen N, Olsen TS, Benediktsson H, Racusen L, Solez K. Reproducibility of the banff classification of renal allograft pathology. *Inter Intraobserver Variat Transplant*. 1995;60(10):1083-1089. doi:10.1097/00007890-199511270-00004
16. Furness PN, Taub N. Convergence of European renal transplant pathology assessment procedures (CERTPAP) project. International variation in the interpretation of renal transplant biopsies: report of the CERTPAP project [published correction appears in Kidney Int 2001 Dec;60(6):2429]. *Kidney Int*. 2001;60(5):1998-2012. doi:10.1046/j.1523-1755.2001.00030.x

17. Mengel M, Sis B, Halloran PF. SWOT analysis of banff: strengths, weaknesses, opportunities and threats of the international banff consensus process and classification system for renal allograft pathology. *Am J Transplant*. 2007;7(10):2221-2226. doi:10.1111/j.1600-6143.2007.01924.x
18. Reeve J, Sellarés J, Mengel M, et al. Molecular diagnosis of T cell-mediated rejection in human kidney transplant biopsies. *Am J Transplant*. 2013;13(3):645-655. doi:10.1111/ajt.12079
19. Emwas AH, Saccenti E, Gao X, et al. Recommended strategies for spectral processing and post-processing of 1D ^{1}H-NMR data of biofluids with a particular focus on urine. *Metabolomics*. 2018;14(3):31. doi:10.1007/s11306-018-1321-4
20. Li B, Tang J, Yang Q, et al. Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. *Sci Rep*. 2016;6:38881. doi:10.1038/srep38881
21. di Guida R, Engel J, Allwood JW, et al. Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics*. 2016;12:93. doi:10.1007/s11306-016-1030-9
22. Ren S, Hinzman AA, Kang EL, Szczesniak RD, Lu LJ. Computational and statistical analysis of metabolomics data. *Metabolomics*. 2015;11:1492-1513. doi:10.1007/s11306-015-0823-6
23. Curran-Everett D. Explorations in statistics: the analysis of ratios and normalized data. *Adv Physiol Educ*. 2013;37(3):213-219. doi:10.1152/advan.00053.2013
24. Waikar SS, Sabbisetti VS, Bonventre JV. Normalization of urinary biomarkers to creatinine during changes in glomerular filtration rate. *Kidney Int*. 2010;78(5):486-494. doi:10.1038/ki.2010.165
25. Goldstein SL. Urinary kidney injury biomarkers and urine creatinine normalization: a false premise or not? *Kidney Int*. 2010;78(5):433-435. doi:10.1038/ki.2010.200
26. Transplant Manitoba. Pediatric kidney program. http://www.transplantmanitoba.ca/transplant-program/pediatric-kidney-transplant. Accessed 24 June 2019.
27. Solez K, Colvin RB, Racusen LC, et al. Banff 07 classification of renal allograft pathology: updates and future directions. *Am J Transplant*. 2008;8(4):753-760. doi:10.1111/j.1600-6143.2008.02159.x
28. Hardinger KL, Koch MJ, Bohl DJ, Storch GA, Brennan DC. BK-virus and the impact of pre-emptive immunosuppression reduction: 5-year results. *Am J Transplant*. 2010;10(2):407-415. doi:10.1111/j.1600-6143.2009.02952.x
29. Merscher S, Fornoni A. Podocyte pathology and nephropathy - sphingolipids in glomerular diseases. *Front Endocrinol (Lausanne)*. 2014;5:127. doi:10.3389/fendo.2014.00127
30. Lam CW, Law CY, Sze KH, To KK. Quantitative metabolomics of urine for rapid etiological diagnosis of urinary tract infection: evaluation of a microbial-mammalian co-metabolite as a diagnostic biomarker. *Clin Chim Acta*. 2015;438:24-28. doi:10.1016/j.cca.2014.07.038
31. Bouatra S, Aziat F, Mandal R, et al. The human urine metabolome. *PLoS One*. 2013;8(9):e73076. doi:10.1371/journal.pone.0073076
32. R Core Team. *R: A Language and Environment for Statistical Computing*. 2018. R Foundation for Statistical Computing, Vienna, Austria: https://www.R-project.org/
33. Nankivell BJ, Agrawal N, Sharma A, et al. The clinical and pathological significance of borderline T cell-mediated rejection. *Am J Transplant*. 2019;19(5):1452-1463. doi:10.1111/ajt.15197
34. Becker JU, Chang A, Nickeleit V, Randhawa P, Roufosse C. Banff borderline changes suspicious for acute T cell-mediated rejection: where do we stand? *Am J Transplant*. 2016;16(9):2654-2660. doi:10.1111/ajt.13784
35. Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc Ser B*. 1996;58(1):267-288. doi:10.1111/j.2517-6161.1996.tb02080.x
36. Hoerl AE. Application of ridge analysis to regression problems. *Chem Eng Prog*. 1962;58(3):54-59.
37. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B*. 2005;67(2):301-320. doi:10.1111/j.1467-9868.2005.00503.x
38. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst*. 2001;58(2):109-130. doi:10.1016/S0169-7439(01)00155-1
39. Worley B, Powers R. Multivariate analysis in metabolomics. *Curr Metabolomics*. 2013;1(1):92-107. doi:10.2174/2213235X11301010092
40. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324
41. Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers. Proceedings of the 5th Annual Workshop of Computational Learning Theory; Vol. 5, 1992:144-152; ACM, Pittsburgh.
42. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20:273-297. doi:10.1023/A:1022627411411
43. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016:785-794; ACM, New York, NY. doi:10.1145/2939672.2939785
44. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw*. 2008;28(5):1-26. doi:10.18637/jss.v028.i05
45. Irizarry RA, Warren D, Spencer F, et al. Multiple-laboratory comparison of microarray platforms [published correction appears in Nat Methods. 2005 Jun;2(6):477]. *Nat Methods*. 2005;2(5):345-350. doi:10.1038/nmeth756
46. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods*. 2018;15(4):255-261. doi:10.1038/nmeth.4612
47. Jiang R, Li W, Li J. mbImpute: an accurate and robust imputation method for microbiome data. *Genome Biol*. 2021;22:192. doi:10.1186/s13059-021-02400-4
48. Kaul A, Mandal S, Davidov O, Peddada SD. Analysis of microbiome data in the presence of excess zeros. *Front Microbiol*. 2017;8:2114.
49. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat Commun*. 2018;9(1):1-9.

50. Voinescu GC, Shoemaker M, Moore H, Khanna R, Nolph KD. The relationship between urine osmolality and specific gravity. *Am J Med Sci*. 2002;323(1):39-42. doi:10.1097/00000441-200201000-00007
51. Chong I, Jun C. Performance of some variable selection methods when multicollinearity is present. *Chemom Intell Lab Syst*. 2005;78(1):103-112. doi:10.1016/j.chemolab.2004.12.011
52. Amaratunga D, Cabrera J, Lee YS. Enriched random forests. *Bioinformatics*. 2008;24(18):2010-2014. doi:10.1093/bioinformatics/btn356
53. Xu B, Huang JZ, Williams G, Wang Q, Ye Y. Classifying very high-dimensional data with random forests built from small subspaces. *Int J Data Warehous Min*. 2012;8(2):44-63. doi:10.4016/jdwm.2012040103
54. Donick D, Lera SC. Uncovering feature interdependencies in high-noise environments with stepwise lookahead decision forests; 2009. https://arxiv.org/pdf/2009.14572.pdf.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Krstic N, Multani K, Wishart DS, Blydt-Hansen T, Cohen Freue GV. The impact of methodological choices when developing predictive models using urinary metabolite data. *Statistics in Medicine*. 2022;41(18):3511-3526. doi: 10.1002/sim.9431